More Efficient Algorithms and Analyses for Unequal Letter Cost Prefix-Free Coding

Mordecai Golin　　　　　Li Jian

Hong Kong UST　　　　Fudan University

golin@cs.ust.hk　　　*lijian83@fudan.edu.cn*

February 1, 2008

Abstract

There is a large literature devoted to the problem of finding an optimal (min-cost) prefix-free code with an unequal letter-cost encoding alphabet of size. While there is no known polynomial time algorithm for solving it optimally there are many good heuristics that all provide additive errors to optimal. The additive error in these algorithms usually depends linearly upon the largest encoding letter size.

This paper was motivated by the problem of finding optimal codes when the encoding alphabet is infinite. Because the largest letter cost is infinite, the previous analyses could give infinite error bounds. We provide a new algorithm that works with infinite encoding alphabets. When restricted to the finite alphabet case, our algorithm often provides better error bounds than the best previous ones known.

Keywords: Prefix-Free Codes. Source-Coding. Redundancy. Entropy.

1   Introduction

Let $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_t\}$ be an *encoding alphabet*. Word $w \in \Sigma^*$ is a *prefix* of word $w' \in \Sigma^*$ if $w' = wu$ where $u \in \Sigma^*$ is a non-empty word. A *Code* over Σ is a collection of words $C = \{w_1, \ldots, w_n\}$. Code C is *prefix-free* if for all $i \neq j$ w_i is not a prefix of w_j. See Figure 1.

Let $cost(w)$ be the *length* or number of characters in w. Given a set of associated probabilities $p_1, p_2, \ldots, p_n \geq 0$, $\sum_i p_i = 1$, the cost of the code is $Cost(C) = \sum_{i=1}^{n} cost(w_i)p_i$. The *prefix coding* problem, sometimes known as the *Huffman encoding* problem is to find a prefix-free code over Σ of minimum cost. This problem is very well studied and has a well-known

x	aaa	aab	ab	b
$cost(x)$	3	5	4	3

x	aaa	aab	ab	$aaba$
$cost(x)$	3	5	4	6

Figure 1: In this example $\Sigma = \{a, b\}$. The code on the left is $\{aaa, aab, ab, b\}$ which is prefix free. The code on the right is $\{aaa, aab, ab, aaba\}$ which is not prefix-free because aab is a prefix of $aaba$. The second row of the tables contain the costs of the codewords when $cost(a) = 1$ and $cost(b) = 3$.
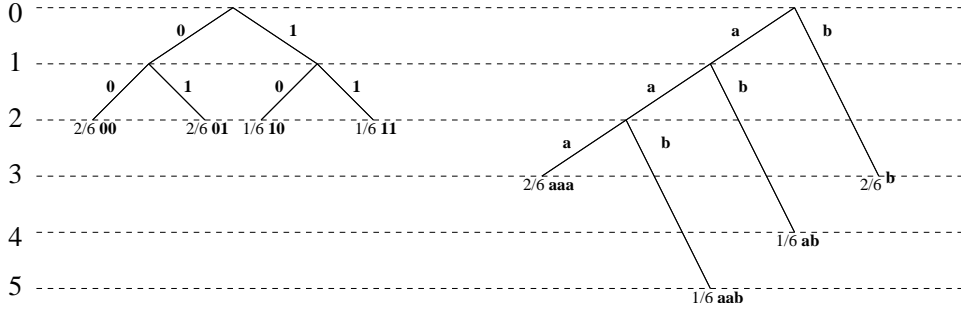

Figure 2: Two min-cost prefix free codes for probabilities $2/6, 2/6, 1/6, 1/6$ and their tree representations. The code on the left is optimal for $c_1 = c_2 = 1$ while the code on the right, the prefix-free code from Figure 1, is optimal for $c_1 = 1, c_2 = 3$.

$O(tn \log n)$-time greedy-algorithm due to Huffman [14] ($O(tn)$-time if the p_i are sorted in non-decreasing order).

Alphabetic coding is the same problem with the additional constraint that the codewords must be chosen in increasing alphabetic order (with respect to the words to be encoded). This corresponds, for example, to the problem of constructing optimal (with respect to average search time) search trees for items with the given access probabilities or frequencies. Such a code can be constructed in $O(tn^3)$ time [16].

One well studied generalization of the problem is to let the encoding letters have different costs. That is, let $\sigma_i \in \Sigma$ have associated cost c_i. The cost of codeword $w = \sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_l}$ will be $cost(w) = \sum_{k=1}^{l} c_{i_k}$, i.e., the sum of the costs of its letters (rather than the length of the codeword) with the cost of the code still being defined as $Cost(C) = \sum_{i=1}^{n} cost(w_i)p_i$ with this new cost function.

The existing, large, literature on the problem of finding a minimal-cost prefix free code when the c_i are no longer equal, which will be surveyed

below, assumes that Σ is a finite alphabet, i.e., that $t = |\Sigma| < \infty$. The original motivation of this paper was to address the problem when Σ is *unbounded*. which, as will briefly be described in Section 3 models certain types of language restrictions on prefix free codes and the imposition of different cost metrics on search trees. The tools developed, though, turn out to provide improved approximation bounds for many of the finite cases as well. More specifically, it was known $[20, 23]$[1] that $\frac{1}{c}H(p_1, \ldots, p_n) \leq OPT$ where $H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$ is the *entropy* of the distribution, c is the unique positive root of the *characteristic equation* $1 = \sum_{i=1}^{t} 2^{-cc_i}$ and OPT is the minimum cost of any prefix free code for those p_i. Note that in this paper, $\log x$ will always denote $\log_2 x$.

The known efficient algorithms create a code T that satisfies

$$C(T) \leq \frac{1}{c}H(p_1, \ldots, p_n) + f(\mathcal{C}) \tag{1}$$

where $C(T)$ is the cost of code T, $\mathcal{C} = (c_1, c_2, \cdots, c_t)$ and $f(\mathcal{C})$ is some function of the letter costs $\mathcal{C}$, with the actual value of $f(\mathcal{C})$ depending upon the particular algorithm. Since $\frac{1}{c}H(p_1, \ldots, p_n) \leq OPT$, code T has an *additive error* at most $f(\mathcal{C})$ from OPT. The $f(\mathcal{C})$ corresponding to the different algorithms shared an almost linear dependence upon the value $c_t = \max(\mathcal{C})$, the largest letter cost. They therefore can not be used for infinite $\mathcal{C}$. In this paper we present a new algorithmic variation (all algorithms for this problem start with the same splitting procedure so they are all, in some sense, variations of each other) with a new analysis:

- (Theorems 2 and 3) For finite $\mathcal{C}$ we derive new additive error bounds $f(\mathcal{C})$ which in many cases, are much better than the old ones.

- (Lemma 9) If $\mathcal{C}$ is infinite but $d_j = |\{m \mid j \leq c_m < j + 1\}|$ is bounded, then we can still give a bound of type (1). For example, if $c_m = 1 + \lfloor \frac{m-1}{2} \rfloor$, i.e., exactly two letters each of length $i, = 1, 2, 3, \ldots$, then we can show that $f(\mathcal{C}) \leq 1 + \frac{3}{\log 3}$.

- (Theorem 4) If $\mathcal{C}$ is infinite but d_i is unbounded then we can not provide a bound of type (1) but, as long as $\sum_{i=1}^{\infty} c_m 2^{-cc_m} < \infty$, we can show that

$$\forall \epsilon > 0, \quad C(T) \leq (1 + \epsilon)\frac{1}{c}H(p_1, \ldots, p_n) + f(\mathcal{C}, \epsilon) \tag{2}$$

---

[1]Note that if $t = 2$ with $c_1 = c_2 = 1$ then $c = 1$ and this reduces to the standard entropy lower bound for prefix-free coding. Although the general lower bound is usually only explicitly derived for finite t, Krause [20] showed how to extend it to infinite t in cases where a positive root of $1 = \sum_{i=1}^{\infty} 2^{-cc_i}$ exists.

where $f(\mathcal{C}, \epsilon)$ is some constant based only on $\mathcal{C}$ and ϵ.

We now provide some more history and motivation.

For a simple example, refer to Figure 2. Both codes have minimum cost for the frequencies $(p_1, p_2, p_3, p_4) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right)$ but under different letter costs. The code $\{00, 01, 10, 11\}$ has minimum cost for the standard Huffman problem in which of $\Sigma = \{0, 1\}$ and $c_1 = c_2 = 1$, i.e., the cost of a word is the number of bits it contains. The code $\{aaa, aab, ab, b\}$ has minimum cost for the alphabet $\Sigma = \{a, b\}$ in which the length of an "a" is 1 and the length of a "b" is 3, i.e., $\mathcal{C} = (1, 3)$.

The unequal letter cost coding problem was originally motivated by coding problems in which different characters have different transmission times or storage costs [4, 22, 18, 27, 28]. One example is the telegraph channel [9, 10, 20] in which $\Sigma = \{\cdot, -\}$ and $c_1 = 1, c_2 = 2$, i.e., in which dashes are twice as long as dots. Another is the (a, b) run-length-limited codes used in magnetic and optical storage [15, 11], in which the codewords are binary and constrained so that each **1** must be preceded by at least a, and at most b, **0**'s. (This example can be modeled by the unequal-cost letter problem by using an encoding alphabet of $r = b - a + 1$ characters $\{0^k 1 : k = a, a + 1, \ldots, b\}$ with associated costs $\{c_i = a + i - 1\}$.)

The unequal letter cost *alphabetic* coding problem arises in designing testing procedures in which the time required by a test depends upon the outcome of the test [19, 6.2.2, ex. 33] and has also been studied under the names *dichotomous search* [13] or the *leaky shower* problem [17].

The literature contains many algorithms for the unequal-cost coding problem. Blachman [4], Marcus [22], and (much later) Gilbert [10] give heuristic constructions without analyses of the costs of the codes they produced. Karp gave the first algorithm yielding an exact solution (assuming the letter costs are integers); Karp's algorithm transforms the problem into an integer program and does not run in polynomial time [18]. Later exact algorithms based on dynamic programming were given by Golin and Rote [11] for arbitrary t and a slightly more efficient one by Bradford et. al. [5] for $t = 2$.. These algorithms run in $n^{\theta(c_t)}$ time where c_t is the cost of the largest letter. Despite the extensive literature, there is no known polynomial-time algorithm for the generalized problem, nor is the problem known to be NP-hard. Golin, Kenyon and Young [12] provide a polynomial time approximation scheme (PTAS). Their algorithm is mainly theoretical and not useful in practice. Finally, in contrast to the non-alphabetic case, alphabetic coding has a polynomial-time algorithm $O(tn^3)$ time algorithm [16].

4

Karp's result was followed by many efficient algorithms [20, 8, 7, 23, 2]. As mentioned above, $\frac{1}{c}H(p_1, \ldots, p_n) \leq OPT$; almost[2] all of these algorithms produce codes of cost at most $C(T) \leq \frac{1}{c}H(p_1, \ldots, p_n) + f(\mathcal{C})$ and therefore give solutions within an *additive error* of optimal. An important observation is that the additive error in these papers $f(\mathcal{C})$ somehow incorporate the cost of the largest letter $c_t = \max(\mathcal{C})$. Typical in this regard is Mehlhorn's algorithm [23] which provides a bound of

$$cC(T) - H(p_1, \ldots, p_n) \leq (1 - p_1 - p_n) + cc_t \qquad (3)$$

Thus, none of the algorithms described can be used to address infinite alphabets with unbounded letter costs.

The algorithms all work by starting with the probabilities in some given order, grouping consecutive probabilities together according to some rule, assigning the same initial codeword prefix to all of the probabilities in the same group and then recursing. They therefore actually create alphabetic codes. Another unstated assumption in those papers (related to their definition of alphabetic coding) is that the order of the c_m is given and must be maintained.

In this paper we are only interested in the general coding problem and not the alphabetic one and will therefore have freedom to dictate the original order in which the p_i are given and the ordering of the c_m. We will actually always assume that $p_1 \geq p_2 \geq p_2 \geq \cdots$ and $c_1 \leq c_2 \leq c_3 \leq \cdots$. These assumptions are the starting point that will permit us to derive better bounds. Furthermore, for simplicity, we will always assume that $c_1 = 1$. If not, we can always force this by uniformly scaling all of the c_i.

For further references on Huffman coding with unequal letter costs, see Abrahams' survey on source coding [1, Section 2.7], which contains a section on the problem.

2 Notations and definitions

There is a very standard correspondence between prefix-free codes over alphabet Σ and $|\Sigma|$-ary trees in which the mth child of node v is labelled with character $\sigma_m \in \Sigma$. A path from the root in a tree to a leaf will correspond to the word constructed by reading the edge labels while walking the path. The

---

[2]As mentioned by Mehlhorn [23], the result of Cot [7] is a bit different. It's a redundancy bound and not clear how to efficiently implement as an algorithm. Also, the redundancy bound is in a very different form involving taking the ratio of roots of multiple equations that makes it difficult to compare to the others in the literature.

tree T corresponding to code $C = \{w_1, \ldots, w_n\}$ will be the tree containing the paths corresponding to the respected words. Note that the leaves in the tree will then correspond to codewords while internal nodes will correspond to prefixes of codewords. See Figures 2 and 5.

Because this correspondence is 1-1 we will speak about codes and trees interchangeably, with the cost of a tree being the cost of the associate code.

Definition 1 *Let C be a prefix free code over Σ and T its associated tree. N_T will denote the set of* internal nodes *of T.*

Definition 2 *Set c to be the unique positive solution to $1 = \sum_{i=1}^{t} 2^{-cc_i}$. Note that if $t < \infty$, then c must exists while if $t = \infty$, c might not exist. We only define c for the cases in which it exists. c is sometimes called the* root *of the characteristic equation* of the letter costs.

Definition 3 *Given letter costs c_i and their associated characteristic root c, let T be a code with those letter costs. If $p_1, p_2, \ldots, p_n \geq 0$ is a probability distribution then the* redundancy *of T relative to the p_i is*

$$R(T; p_1, \ldots, p_n) = C(T) - \frac{1}{c} H(p_1, \ldots, p_n).$$

We will also define the normalized redundancy *to be*

$$\mathrm{NR}(T; p_1, \ldots, p_n) = cR = cC(T) - H(p_1, \ldots, p_n).$$

If the p_i and T are understood, we will write $R(T)$ ($\mathrm{NR}(T)$) or even R (NR).

We note that many of the previous results in the literature, e.g., (3) from [23], were stated in terms of NR. We will see later that this is a very natural measure for deriving bounds. Also, note that by the lower bound previously mentioned, $C(T) \geq \frac{1}{c} H(p_1, \ldots, p_n)$ for all T and p_i, so $R(T; p_1, \ldots, p_n)$ is a good measure of absolute error.

3 Examples of Unequal-Cost Letters

It is very easy to understand the unequal-cost letter problem as modelling situations in which different characters have different transmission times or storage costs [4, 22, 18, 27, 28]. Such cases will all have finite alphabets. It is not a-priori as clear why infinite alphabets would be interesting. We now discuss some motivation.

In what follows we will need some basic language notation. A language $\mathcal{L}$ is just a set of words over alphabet Σ. The *concatenation* of languages A and B is $AB = \{ab \mid a \in A, b \in B\}$. The i-fold concatenation, $\mathcal{L}^i$, is defined by $\mathcal{L}^0 = \{\lambda\}$ (the language containing just the empty string), $\mathcal{L}^1 = \mathcal{L}$ and $\mathcal{L}^i = \mathcal{L}\mathcal{L}^{i-1}$. The *Kleene star* of $\mathcal{L}$, is $\mathcal{L} = \bigcup_{i=0}^{\infty} \mathcal{L}^i$.

We start with cost vector $\mathcal{C} = \{1, 2, 3, \dots, \}$ i.e, $\forall m > 0, c_m = m$. An early use of this problem was [24]. The idea there was to construct a tree (not a code) in which the internal pointers to children were stored in a linked list. Taking the mth pointer corresponds to using character σ_m. The time that it takes to *find* the mth pointer is proportional to the location of the pointer in the list. Thus (after normalizing time units) $c_m = m$.

We now consider a generalization of the problem of **1**-ended codes. The problem of finding min-cost prefix free code with the additional restriction that all codewords end with a 1 was studied in [3, 6] with the motivation of designing self-synchronizing codes. One can model this problem as follows. Let $\mathcal{L}$ be a language. In our problem,

$$\mathcal{L} = \{w \in \{0,1\}^* \mid \text{the last letter in } w \text{ is a } \mathbf{1}\}.$$

We say that a code C is in $\mathcal{L}$ if $C \subseteq \mathcal{L}$. The problem is to find a minimum cost code among all codes in $\mathcal{L}$.

Now suppose further that $\mathcal{L}$ has the special property that $\mathcal{L} = \mathcal{Q}^*$ where $\mathcal{Q}$ is itself a prefix-free language. Then every word in $\mathcal{L}$ can be uniquely decomposed as the concatenation of words in $\mathcal{Q}$. If the decomposition of $w \in \mathcal{L}$ is $w = q_1 q_2 \dots q_r$ for $q_i \in \mathcal{Q}$ then $cost(w) = \sum_{i=1}^{r} cost(q_i)$. We can therefore model the problem of finding a minimum cost code among all codes in $\mathcal{L}$ by first creating an infinite alphabet $\Sigma_{\mathcal{Q}} = \{\sigma_q \mid q \in Q\}$ with associated cost vector $C_{\mathcal{Q}}$ (in which the length of σ_q is $cost(q)$) and then solving the minimal cost coding problem for $\Sigma_{\mathcal{Q}}$ with those associated costs. For the example of **1**-ended codes we set $\mathcal{Q} = \{1, 01, 001, 0001, \dots\}$ and thus have $\mathcal{C} = \{1, 2, 3, \dots, \}$ i.e, an infinite alphabet with $c_m = 1$ for all $m \geq 1$.

Now consider generalizing the problem as follows. Suppose we are given an unequal cost coding problem with *finite* alphabet $\Sigma = \{\sigma_1, \dots, \sigma_t\}$ and associated cost vector $\mathcal{C} = (c_1, \dots, c_t)$. Now let $\Sigma' \subset \Sigma$ and define

$$\mathcal{L} = \Sigma^* \Sigma' = \{w \in \Sigma^* \mid \text{the last letter in } w \text{ is in } \Sigma'\}.$$

Now note that $\mathcal{L} = D^*$ where $D = (\Sigma - \Sigma')^* \Sigma'$ is a prefix-free language. We can therefore model the problem of finding a minimum cost code among all codes in $\mathcal{L}$ by solving an unequal cost coding problem with alphabet Σ_D

and $\mathcal{C}_D$. The important observation is that

$$d_j = |\{d \in \Sigma_D \mid cost(d) = j\}|,$$

the number of letters in Σ_D of length j, satisfies a linear recurrence relation. Bounding redundancies for these types of $\mathcal{C}$ will be discussed in Section 6, Case 4.

As an illustration, consider $\Sigma = \{1, 2, 3\}$ with $\mathcal{C} = (1, 1, 2)$ and $\Sigma' = \{1\}$; our problem is find minimal cost prefix free codes in which all words end with a 1. $\mathcal{L} = \{1, 2, 3\}^*\{1\} = D^*$, where $D = \{2, 3\}^*\{1\}$. The number of characters in Σ_D with length j is

$$d_1 = 1, \ d_2 = 1, \ d_3 = 2, \ d_4 = 3, \ d_5 = 5,$$

and, in general, $d_{i+2} = d_{i+1} + d_i$, so $d_i = F_i$, the Fibonacci numbers.

We conclude with a very natural $\mathcal{L}$ for which we do *not* know how to analyze the redundancy. In Section 6, Case 5 we will discuss why this is difficult.

Let $\mathcal{L}$ be the set of all "balanced" binary words[3], i.e., all words which contain exactly as many **0**'s as **1**'s. Note that $\mathcal{L} = \mathcal{D}^*$ where $\mathcal{D}$ is the set of all non-empty balanced words w such that no prefix of w is balanced. Let

$$\ell_j = |\{w \in \mathcal{L} \mid cost(w) = j\}|, \quad d_j = |\{w \in D \mid cost(w) = j\}|$$

and set $L(z) = \sum_{n=0}^{\infty} \ell_n z^n$ and $D(z) = \sum_{n=0}^{\infty} d_n z^n$ to be their associated generating functions. If $\mathcal{L} = D^*$, then standard generating function rules, see e.g., [25], state that $L(z) = (1 - D(z))^{-1}$. Observe that $l_n = 0$ if n is odd and $l_n = \binom{n}{n/2}$ for n even, so

$$L(z) = \sum_{n=0}^{\infty} \binom{2n}{n} (z^2)^n = \frac{1}{\sqrt{1 - 4z^2}}$$

and

$$\sum_{n=1}^{\infty} d_j z^j = D(z) = 1 - \sqrt{1 - 4z^2}.$$

This can then be solved to see that, for even $n > 0$, $d_n = 2C_{n/2-1}$ where $C_i = \frac{1}{i+1}\binom{2i}{i}$ is the ith Catalan number. For $n = 0$ or odd n, $d_n = 0$.
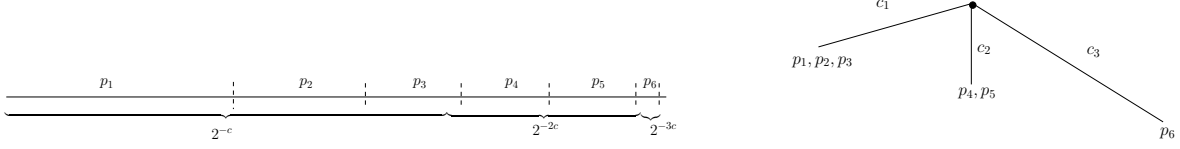
Figure 3: The first splitting step for a case when $n = 6$ $c_1 = 1$, $c_2 = 2$, $c_3 = 3$ and the associated preliminary tree. This step groups p_1, p_2, p_3 as the first group, p_3, p_4 as the second and p_5 by itself. Note that we haven't yet formally explained yet *why* we've grouped the items this way.
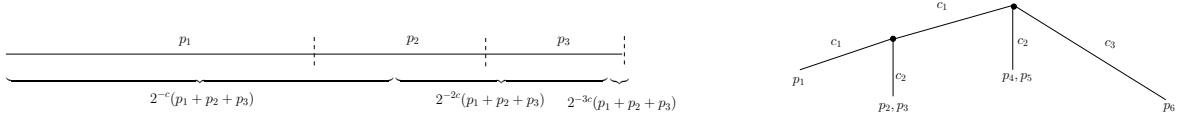

Figure 4: In the second split, p_1 is kept by itself and p_2, p_3 are grouped together.
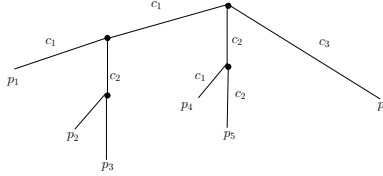

Figure 5: After two more splits, the final coding tree is constructed. The associated code is $\{\sigma_1\sigma_1, \sigma_1\sigma_2\sigma_1, \sigma_1\sigma_2\sigma_2, \sigma_2\sigma_1, \sigma_2\sigma_2, \sigma_3\}$

4   The algorithm

All of the provably efficient heuristics for the problem, e.g., [20, 8, 7, 23, 2], use the same basic approach, which itself is a generalization of Shannon's original binary splitting algorithm [26]. The idea is to create t *bins*, where bin m has weight 2^{-cc_m} (so the sum of all bin weights is 1). The algorithms then try to partition the probabilities into the bins; bin m will contain a set of contiguous probabilities $p_{l_m}, p_{l_m+1}, \ldots, p_{r_m}$ whose sum will have

---

[3]This also generalizes a problem from [21] which provides heuristics for constructing a min-cost prefix-free code in which the expected number of **0**'s equals the expected number of **1**'s.

total weight "close" to 2^{-cc_m}. The algorithms fix the first letter of all the codewords associated with the p_k in bin m to be σ_m. After fixing the first letter, the algorithms then recurse, normalizing $p_{l_m}, p_{l_m+1}, \ldots, p_{r_m}$ to sum to 1, taking them as input and starting anew. The various algorithms differ in how they group the probabilities and how they recurse. See Figures 3, 4 and 5 for an illustration of this generic procedure.

Here we use a generalization of the version introduced in [23]. The algorithm first preprocesses the input and calculates all $P_k = p_1 + p_2 + \ldots + p_k$ ($P_0 = 0$) and $s_k = p_1 + p_2 + \ldots + p_{k-1} + \frac{p_k}{2}$. Note that if we lay out the p_i along the unit interval in order, then s_k can be seen as the *midpoint* of interval p_i. It then partitions the probabilities into ranges, and for each range it constructs left and right boundaries L_m, R_m. p_k will be assigned to bin m if it "falls" into the "range" $[L_m, R_m)$.

If the interval p_k falls into the range, i.e., $L_m \leq P_{k-1} < P_k < R_m$ then p_k should definitely be in bin m. But what if p_k spans two (or more) ranges, e.g., $L_m \leq P_{k-1} < R_m < P_k$? To which bin should p_k be assigned? The choice made by [23] is that p_k is assigned to bin m if $s_k = p_1 + p_2 + \ldots + p_k/2$ falls into $[L_m, R_m)$, i.e., the midpoint of p_k falls into the range.

Our procedure $CODE(l, r, U)$ will build a prefix-free code for $p_l, \ldots, p_r$ in which every code word starts with prefix U. To build the entire code we call $CODE(1, n, \lambda)$, where λ is the empty string.

The procedure works as follows (Figure 6 gives pseudocode and Figures 7, 8 and 9 illustrate the concepts):

Assume that we currently have a prefix of U assigned to $p_l, \ldots, p_r$. Let v be node in the tree associated with U. Let $w(v) = \sum_{k=l}^{r} p_k$.

(i) If $l = r$ then word U is assigned to p_l. Correspondingly, v is a leaf in the tree with weight $w(v) = p_l$.

(ii) Otherwise let $L = P_l$ and $R = P_r$. Split $R - L = w(v)$ into t ranges[4] as follows.

$$\forall 1 \leq m \leq t, \quad L_m = L + (R - L) \sum_{i=1}^{m-1} 2^{-cc_i}, \quad R_m = L + (R - L) \sum_{i=1}^{m} 2^{-cc_i}.$$

Insert p_k, $l \leq k \leq r$ in bin m if $s_k \in [L_m, R_m)$. Bin m will thus contain the p_k in $I_m^*(v) = \{k \mid L_m \leq s_k < R_m\}$.

We now shift the items p_k *leftward* in the bins as follows. Walk through the bins from left to right. If the current bin already contains some p_k, continue to the next bin. If the current bin is empty, take the first p_k that appears in a bin to the right of the current one, shift p_k into the current

---

[4]In the description, t is permitted to be finite or infinite.

$CODE(l, r, U)$;
{Constructs codewords $U_l, U_{l+1}, \ldots, U_r$ for $p_l, p_{l+1}, \ldots, p_r$.
 U is previously constructed common prefix of $U_l, U_{l+1}, \ldots, U_r$.}
If $l = r$
    then codeword U_l is set to be U.
else   *{Distribute p_is into initial bins I_m^*}*
    $L = P_{l-1}$; $R = P_r$; $w = R - L$
    $\forall m$, let $L_m = L + w \sum_{i=1}^{m-1} 2^{-cc_i}$ and $R_m = L_m + w2^{-cc_m}$.
        set $I_m^* = \{k \mid L_m \le s_k < R_m\}$ }

    {Shift the bins to become final I_m. Afterwards,
    all bins $> M$ are empty, all bins $\le M$ non-empty
    and $\forall m \le M$, $I_m = \{l_m, \ldots r_m\}$}}

    {shift left so there are no empty "middle" bins.}
    $M = 0$; $k = l$;
    while $k \le r$ **do**
        $M = M + 1$;
        $l_M = k$; $r_M = \max\Big(\{k\}\bigcup\{i > k \mid i \in I_M^*\}\Big)$;
        $k = r_M + 1$;

    {If all p_i's are in first bin, shift p_r to 2nd bin }
    if $r_1 = r$ **then**
        $M = 2$;
        $r_1 = r - 1$; $l_2 = r_2 = r$;

    for $m = 1$ **to** M **do**
        $CODE(l_m, r_m, U\sigma_m)$;

Figure 6: Our algorithm. Note that the first step of creating the I_m^* was written to simplify the development of the analysis. In practice, it is not needed since I_m^* is only used to find $\max\{i > k \mid i \in I_m^*\}$ and this value can be calculated using binary search at the time it is required.
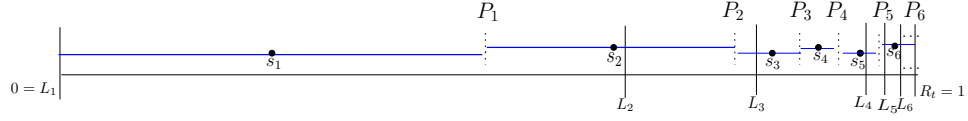
Figure 7: The first step in our algorithm's splitting procedure. $n = 6$. $L_i = \sum_{m=1}^{i-1} 2^{-cc_m}$. Note that even though only the first 5 L_i are shown, there might be an infinite number of them (if $t = \infty$). Note too that, for $0 < i$, $L_i = R_{i-1}$.
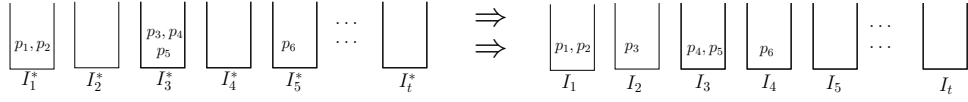

Figure 8: The splitting procedure performed on the above example creates the bins I_m^* on the left. The shifting procedure then creates the I_m on the right.
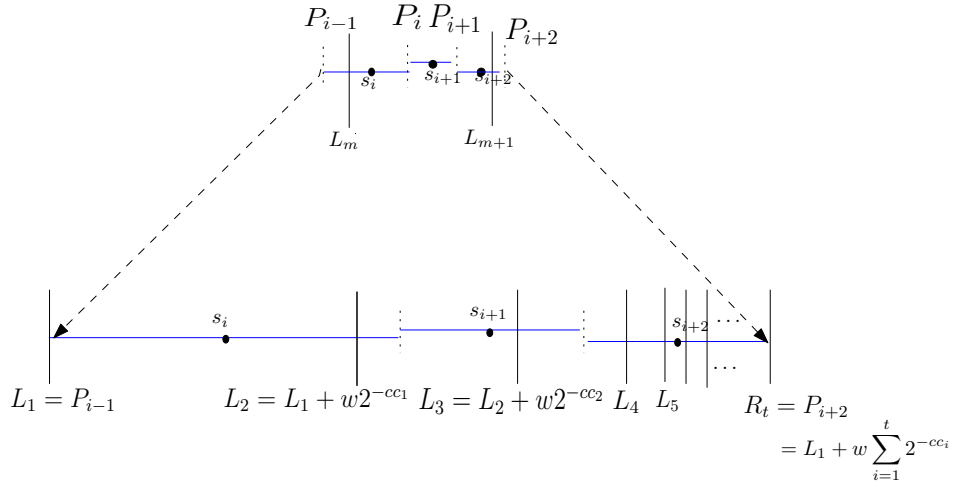

Figure 9: An illustration of the recursive step of the algorithm. p_i, p_{i+1}, p_{i+2} have been grouped together. In the next splitting step, the interval operated on has length $w = p_i + p_{i+1} + p_{i+2}$.

12

bin and walk to the next bin. Stop when all p_k have been seen. Let $I_m(v)$ denote the items in the bins after this shifting.

Note that after performing this shifting there is some $M(v)$ such that all bins $m \leq M(v)$ are nonempty and all bins $m > M(v)$ are empty. Also notice that it is not necessary to actually construct the $I_m^*(v)$ first. We only did so because they will be useful in our later analysis. We can more efficiently construct the $I_m(v)$ from scratch by walking from left to right, using a binary search each time, to find the rightmost item that should be in the current bin. This will take $O(M(v) \log(l - r))$ time in total.

We then check if all of the items are in $I_1(v)$. If they are, we take p_r and move it into $I_2(v)$ (and set $M(v) = 2$).

Finally, after creating the all of the $I_m(v)$ we let $l_m = \min\{k \in I_m(v)\}$ and $r_m = \max\{k \in I_m(v)\}$ and recurse, for each $m < M(v)$ building $CODE(l_m, r_m, U\sigma_m)$

It is clear that the algorithm builds some prefix code with associated tree T. As defined, let N_T be the set of internal nodes of T. Since every internal node of T has at least two children, $\sum_{v \in N_T} M(v) \leq 2n - 1$.

The algorithm uses $O(1)$ time at each of its n leaves and $O(M(v) \log n)$ time at node v. Its total running time is thus bounded by

$$n + \sum_{v \in N_T} \log n M(v) = O(n \log n)$$

with no dependence upon t.

For comparison, we point out the algorithm in [23] also starts by first finding the $I_m^*(v)$. Since it assumed $t < \infty$, its shifting stage was much simpler, though. It just shifted p_l into the first bin and p_r into the tth bin (if they were not already there).

We will now see that our modified shifting procedure not only permits a finite algorithm for infinite encoding alphabets, but also often provides a provably better approximation for *finite* encoding alphabets.

5 Analysis

In the analysis we define $w_m^*(v) = \sum_{k \in I_m^*(v)} p_k$, $w_m(v) = \sum_{k \in I_m(v)} p_k$. Note that $w(v) = \sum_{m=1}^t w_m^*(v) = \sum_{m=1}^t w_m(v) = \sum_{k=l}^r p_k$.

We first need three Lemmas from [23]. The first was proven by recursion on the nodes of a tree, the second followed from the definition of the splitting procedure and the third from the second by some algebraic manipulations.

Lemma 1 *[23] Let T be a code tree and N_T be the set of all internal nodes of T. Then*

1. *The cost $C(T)$ of the code tree T is*

$$C(T) = \sum_{v \in N_T} \sum_{m=1}^{t} c_m \cdot w_m(v)$$

2. *The entropy $H(p_1, p_2, \ldots, p_n)$ is*

$$H(p_1, p_2, \ldots, p_n) = \sum_{v \in N_T} w(v) \cdot H\left(\frac{w_1(v)}{w(v)}, \frac{w_2(v)}{w(v)}, \ldots\right)$$

Lemma 1 permits expressing the normalized redundancy of T as

$$
\begin{aligned}
NR(T) &= c \cdot C(T) - H(p_1, p_2, \ldots, p_n) \\
&= \sum_{v \in N_T} w(v) \left[\sum_{m=1}^{t} \frac{w_m(v)}{w(v)} \left(\log 2^{cc_m} + \log \frac{w_m(v)}{w(v)} \right) \right].
\end{aligned}
$$

Set

$$E(v, m) = \frac{w_m(v)}{w(v)} \left(\log 2^{cc_m} + \log \frac{w_m(v)}{w(v)} \right).$$

Note that $NR(T) = \sum_{v \in N_T} w(v) \left(\sum_{m=1}^{t} E(v, m) \right)$. For convenience we will also define

$$
\begin{aligned}
E^*(v, m) &= \frac{w_m^*(v)}{w(v)} \left(\log 2^{cc_m} + \log \frac{w_m^*(v)}{w(v)} \right), \\
\mathrm{NR}^*(T) &= \sum_{v \in N_T} w(v) \left(\sum_{m=1}^{t} E^*(v, m) \right)
\end{aligned}
$$

The analysis proceeds by bounding the values of $\mathrm{NR}^*(T)$ and $\mathrm{NR}(T) - \mathrm{NR}^*(T)$.

Lemma 2 *[23][5] (note: In this Lemma, the p_i can be arbitrarily ordered.) Consider any call $CODE(l, r, U)$ with $l < r$. Let node v correspond to the word U. Let sets $I_1^*, I_2^*, \ldots$ be defined as in procedure CODE.*

a) *If $I_m^* = \emptyset$, then $w_m^*(v) = 0$.*

---

[5]slightly rewritten for our notation

b) If $I_m^* = \{e\}$. then $w_m^*(v) = p_e$.

c) If $|I_m^*| \geq 2$. Let $e = \min I_m^*$ and $f = \max I_m^*$.
   If $m = 1$, then $\frac{w_m^*(v)}{w(v)} \leq 2^{-cc_1} + \frac{p_f}{2w(v)} \leq 2 \cdot 2^{-cc_1}$.
   If $m = t$ (note that this case requires $t < \infty$) then

$$\frac{w_t^*(v)}{w(v)} \leq 2^{-cc_t} + \frac{p_e}{2w(v)} \leq 2 \cdot 2^{-cc_t}$$

   If $2 \leq m < t$, then $\frac{w_m^*(v)}{w(v)} \leq 2^{-cc_m} + \frac{p_e + p_f}{2w(v)} \leq 2 \cdot 2^{-cc_m}$

Lemma 3 *[23]* (note: In this Lemma, the p_i can be arbitrarily ordered.)
In case (c) of Lemma 2, $E^*(v, m) \leq \frac{p_e + p_f}{w(v)}$.
Furthermore, if $m = 1$ then $E^*(v, m) \leq \frac{p_f}{w(v)}$,
while if $m = t$, then $E^*(v, m) \leq \frac{p_e}{w(v)}$.

Corollary 4 *If the p_i are sorted in nondecreasing order then in case (c) of Lemma 2,*
   if $m = 1$, $E^*(v, m) \leq \frac{p_f}{w(v)}$, while if $m > 1$, then $E^*(v, m) \leq \frac{2p_e}{w(v)}$.

Lemma 5

$$\mathrm{NR} - \mathrm{NR}^* \leq c(c_2 - c_1) \sum_{i \in A} p_i$$

where

$$A = \{i \mid i \text{ is right shifted by the algorithm at some step}\}.$$

Note: p_1 *can never be right shifted, so* $\sum_{i \in A} p_i \leq 1 - p_1$.

Proof: Define

$$X(v) = \sum_{m=1}^{t} w(v)E(v, m) \quad \text{and} \quad X^*(v) = \sum_{m=1}^{t} w(v)E^*(v, m)$$

Note that $\mathrm{NR} = \sum_{v \in N_T} X(v)$ and $\mathrm{NR}^* = \sum_{v \in N_T} X^*(v)$ For each v we will compare $X^*(v)$ and $X(v)$. If no shifts were performed while processing v, then $X^*(v) = X(v)$ and there is nothing to do. We now examine the two mutually exclusive cases of performing left shifts or performing a right shift.

Left shifts:
Every step in our left-shifting procedure involves taking a probability out of some bin m and and moving it into some currently empty bin $r < m$. Let $w_m'(v)$ be the weight in bin m before that shift and p be the probability of

15

the item being shifted. Note that the original weight of bin r was $w'_r(v) = 0$ while after the shift, bin r will have weight p and bin m weight $w'_m(v) - p$. We use the trivial fact

$$\forall p, q > 0, \quad p \log p + q \log q \leq (p+q) \log(p+q) \tag{4}$$

Setting $q = w'_m(v) - p$ in (4) implies

$$p \log \frac{p}{w(v)} + (w'_m(v) - p) \log \frac{(w'_m(v) - p)}{w(v)} \leq w'_m(v) \log \frac{w'_m(v)}{w(v)}.$$

Furthermore, the fact that the c_i are nonincreasing implies

$$p \log 2^{cc_r} + (w'_m(v) - p) \log 2^{cc_m} \leq w'_m(v) \log 2^{cc_m} \tag{5}$$

Combining the two last equations gives that

$$p \left(\log 2^{cc_r} + \log \frac{p}{w(v)} \right) + (w'_m(v) - p) \left(\log 2^{cc_m} + \log \frac{w'_m(v) - p}{w(v)} \right)$$

is

$$\leq w'_m(v) \left(\log 2^{cc_m} + \log \frac{w'_m(v)}{w(v)} \right)$$

Since moving from $X^*(v)$ to $X(v)$ involves only operations in which probabilities are shifted to the left into an empty bucket, the analysis above implies that $X(v) \leq X^*(v)$.

Right shifts:

Consider node v. Suppose that all of the probabilities in v fall into I_1^* with $I_1^* = \{p_e, \ldots, p_f\}$ and $e \neq f$. Since p_f starts in bin 1, p_e must be totally contained in bin 1, so $p_e \leq 2^{-cc_1} w(v)$. The algorithm shifts p_f to the right giving $I_1 = I_1 - \{p_f\}$ and $I_2 = \{p_f\}$. The p_i are nonincreasing so $p_f \leq p_e$.

$$E(v, 2) = \frac{p_f}{w(v)} \left(\log 2^{cc_2} + \log \frac{p_f}{w(v)} \right) \leq \frac{p_f}{w(v)} (cc_2 - cc_1).$$

Also $E(v, 1) \leq E^*(v, 1)$. Thus

$$
\begin{aligned}
w(v) \sum_{m=1}^{t} E(v, m) &= w(v) E(v, 1) + w(v) E(v, 2) \\
&\leq w(v) E^*(v, 1) + p_f(cc_2 - cc_1)
\end{aligned}
$$

Once a p_f is right-shifted it immediately becomes a leaf and can never be right-shifted again.

Combining the analyses of left shifts and right shifts gives

$$\text{NR} = \sum_{v \in N_T} X(v) \leq \sum_{v \in N_T} X^*(v) + c(c_2 - c_1) \sum_{i \in A} p_i = \text{NR}^* + c(c_2 - c_1) \sum_{i \in A} p_i.$$

$\square$

Lemma 6

$$\text{NR}^* \leq 2(1 - p_1) + \sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)| = 1}} w(v) E^*(v, m).$$

Proof: We evaluate NR^* by partitioning it into

$$\text{NR}^* = \sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)| \geq 2}} w(v) E^*(v, m) + \sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)| = 1}} w(v) E^*(v, m). \quad (6)$$

We use a generalization of an amortization argument developed in [23] to bound the first summand. From Corollary 4 we know that if $|I_m^*(v)| \geq 2$ with $e = \min I_m^*$ and $f = \max I_m^*$ then $w(v) E^*(v, m)$ is at most (a) p_f or (b) $2p_e$, depending upon whether (a) $m = 1$, or (b) $m > 1$.

Suppose that some p_i appears as $2p_i$ in such a bound because $i = \min I_m^*(v)$, i.e., case (b). Then, in all later recursive steps of the algorithm i will always be the leftmost item in bin 1 and will therefore not be used in any later case (a) or (b) bound.

Now suppose that some p_i appears in such a bound because $i = \max I_m^*(v)$, i.e., case (a). Then in all later recursive steps of the algorithm, i will always be the rightmost item in the rightmost non-empty bin. The only possibility for it to be used in a later bound is if becomes the rightmost item in bin 1, i.e., all of the probabilities are in $I_1^*(v)$. In this case, p_i is used for a second case (a) bound. Note that if this happens, then p_i is immediately right shifted, becomes a leaf in bin 2, and is never used in any later recursion.

Any given probability p_i can therefore be used either once as a case (b) bound and contribute $2p_i$ or twice as a case (b) bound and again contribute $2 \cdot p_i$. Furthermore, p_1 can never appear in a case (a) or (b) bound because, until it becomes a leaf, it can only be the leftmost item in bin 1. Thus

$$\sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)| \geq 2}} w(v) E^*(v, m) \leq 2(1 - p_1). \quad (7)$$

$\square$

17

Note: In Melhorn's original proof [23] the value corresponding to the RHS of (7) was $(1 - p_1 - p_n)$. This is because the shifting step of Mehlhorn's algorithm guaranteed that $|I_t^*(v)| \neq 0$ and thus there was a symmetry between the analysis of leftmost and rightmost. In our situation t might be infinity so we can not assume that the rightmost non-empty bin is t and we get $2(1 - p_1)$ instead.

Combining this Lemma with Lemma 5 gives

Corollary 7

$$\text{NR} \leq 2(1 - p_1) + c(c_2 - c_1) \sum_{i \in A} p_i + \sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)|=1}} w(v)E^*(v, m).$$

We will now see different bounds on the last summand in the above expression. Section 6 compares the results we get to previous ones for different classes of $\mathcal{C}$. Before proceeding, we note that any p_i can only appear as $I_m^*(v) = \{p_i\}$ for at most one (m, v) pair. Furthermore, if p_i does appear in such a way, then it can not have been made a leaf by a previous right shift and thus $p_i \notin A$.

We start by noting that, when $t \leq \infty$ our bound is never worse than 1 plus the old bound of $(1 - p_1 - p_n) + cc_t$ stated in (3).

Theorem 1 *If $t < \infty$ then*

$$\text{NR} \leq 2(1 - p_1) + cc_t$$

Proof: If $I_m^*(v) = \{p_i\}$ then $w(v)E^*(v, m) \leq p_i cc_m$ so

$$\sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)|=1}} w(v)E^*(v, m) \leq \sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)|=1}} p_i cc_m \leq cc_t \sum_{i \notin A} p_i.$$

The theorem then follows from Corollary 7. □

For a tighter analysis we will need a better bound for the case $|I_m^*| = 1$.

Lemma 8 *(a) Let $v \in N_T$. Suppose i is such that $i \in I_m^*(v)$. then*

$$\frac{p_i}{w(v)} \leq 2 \cdot \sum_{j=m}^{t} \frac{1}{2^{c \cdot c_j}}$$

(b) Further suppose there is some $m' > m$ such that $I_{m'}^* \neq \emptyset$. Then

$$\frac{p_i}{w(v)} \leq 2 \cdot \sum_{j=m}^{m'} \frac{1}{2^{c \cdot c_j}} \leq 3 \cdot \sum_{j=m}^{m'-1} \frac{1}{2^{c \cdot c_j}}$$

18

Proof: Consider the call $CODE(l, r, U)$ at node v. The fact that $i \in I_m^*(v)$ implies $L + \sum_{j=1}^{m-1} 2^{-cc_j} = L_m \le s_i$. To prove (a) just note that

$$s_i + \frac{p_i}{2} = P_i \le P_r = R = L + w(v) \sum_{i=1}^{t} 2^{-cc_i}.$$

So $\frac{p_i}{2} \le w(v) \sum_{j=m}^{t} \frac{1}{2^{c \cdot c_j}}$.

To prove part (b) let $i' \in I_{m'}^*$. Then

$$s_i + \frac{p_i}{2} = P_i \le s_{i'} < L + w(v) \sum_{j=1}^{m'} 2^{-cc_j}.$$

So $\frac{p_i}{2} \le w(v) \sum_{j=m}^{m'} \frac{1}{2^{c \cdot c_j}}$. The final inequality follows from the fact that $c_{m'-1} \le c_m$. $\qquad\square$

Definition 4 *Set $\beta_m = 2^{cc_m} \sum_{i=m}^{t} 2^{-cc_i}$ and $\beta = \sup\{\beta_m \mid 1 \le m \le t\}$*

We can now prove our first improved bound:

Theorem 2 *If $\beta < \infty$ then*

$$\text{NR} \le 2(1 - p_1) + \max\big(c(c_2 - c_1), 1 + \log \beta\big)$$

Proof: Note that using Definition 4 and Lemma 8(a) we can bound the last summand in Corollary 7 as

$$
\begin{aligned}
w(v)E^*(m, v) &= w_m^*(v)\left(\log 2^{cc_m} + \log \frac{w_m^*(v)}{w(v)}\right) \\
&\le w_m^*(v) \log\left(2^{cc_m} 2 \sum_{i=m}^{t} 2^{-cc_i}\right) \\
&\le w_m^*(v)(1 + \log \beta)
\end{aligned}
$$

If $w_m^*(v) = \{i\}$ then i was not a leaf in any previous step and therefore could not have been right shifted, so $i \notin A$. Thus

$$\sum_{\substack{v \in N_T}} \sum_{\substack{1 \le m \le t \\ |I_m^*(v)|=1}} w(v)E^*(v, m) \le (1 + \log \beta) \sum_{i \notin A} p_i.$$

$\qquad\square$

This immediately gives an improved bound for many finite cases because, if $t < \infty$, then $\beta_m = 2^{cc_m} \sum_{i=m}^{t} 2^{-cc_i} \le t - m + 1$ so $\beta \le t$. Thus

19

Theorem 3 *If t is finite then*

$$NR \leq 2(1 - p_1) + \max\big(c(c_2 - c_1), 1 + \log t\big)$$

Definition 5 *For all $j \geq 1$, set*

$$d_j = |\{i \mid j \leq c_i < j + 1\}|.$$

This permits us to give another general bound that also works for many infinite alphabets.

Lemma 9 *If $d_j = O(1)$, then $NR = O(1)$. In particular, if $\forall j$, $d_j \leq K$ then $\beta \leq \frac{2^c K}{1 - 2^{-c}}$ so, from Theorem 2,*

$$NR \leq 2(1 - p_1) + \max \left(c(c_2 - c_1),\, 1 + c + \log \left(\frac{K}{1 - 2^{-c}} \right) \right).$$

Furthermore, if all of the c_i are integers, then $\beta \leq \frac{K}{1 - 2^{-c}}$ and

$$NR \leq 2(1 - p_1) + \max \left(c(c_2 - c_1),\, 1 + \log \left(\frac{K}{1 - 2^{-c}} \right) \right).$$

Proof: Since $c_1 = 1$ we must have $2^{-c} < 1$. Thus, for all $m \geq 1$, if $\ell \leq c_m < \ell + 1$ then

$$
\begin{aligned}
\beta_m &= 2^{cc_m} \sum_{i=m}^{t} 2^{-cc_m} \\
&\leq 2^{c(\ell+1)} \sum_{j=\ell}^{\infty} d_j 2^{-cj} \\
&\leq 2^c K 2^{c\ell} \sum_{j=\ell}^{\infty} 2^{-cj} = \frac{2^c K}{1 - 2^{-c}}
\end{aligned}
$$

which is independent of m and ℓ. The analysis when the c_i are all integers is similar. $\qquad\square$

For general infinite alphabets we are not able to derive a constant redundancy bound but we can prove

Theorem 4 *If $\mathcal{C}$ is infinite and $\sum_{m=1}^{\infty} c_m 2^{-cc_m} < \infty$, then, for every $\epsilon > 0$*

$$R \leq \epsilon \frac{1}{c} H(p_1, \ldots, p_n) + f(\mathcal{C}, \epsilon) \tag{8}$$

20

where $f(\mathcal{C}, \epsilon)$ is some constant based only on $\mathcal{C}$ and ϵ. Note that this is equivalent to stating that

$$C(T) \leq (1 + \epsilon)OPT + f(\mathcal{C}, \epsilon)$$

Proof: We must bound the

$$\sum_{v \in N_T} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)|=1}} w(v)E^*(v, m)$$

term from the right hand side of Corollary 7. Recall that $|I_m^*(v)| = 1$ means that $\exists i$ such that $I_m^*(v) = \{i\}$, i.e., $w_m^*(v) = p_i$ and thus $w(v)E^*(v, m) \leq p_i cc_m$.

Let N_ϵ be a value to be determined later and $m_\epsilon = \max\{m \mid c_m \leq N_\epsilon\}$. Since no probability appears more than once in the sum we can write

$$\sum_{v \in N_T} \sum_{\substack{1 \leq m \leq m_\epsilon \\ |I_m^*(v)|=1}} w(v)E^*(v, m) \leq cN_\epsilon.$$

To analyze the remaining cases, fix v. Consider the set of indices

$$M_v = \{m \mid (m > m_\epsilon) \text{ and } |I_m^*(v)| = 1|\}.$$

Sort these indices in increasing order so that $M_v = \{m_1, m_2, \ldots, m_r\}$ for some r with $m_1 < m_2 < \cdots < m_r$. Let i_j be such that $I_{m_j}^*(v) = \{p_{i_j}\}$. Thus

$$\sum_{\substack{m_\epsilon < m \\ |I_m^*(v)|=1}} w(v)E^*(v, m) = \sum_{j=1}^{r} w(v)E^*(v, m_j) \leq \sum_{j=1}^{r} p_{i_j} cc_{m_j}$$

Lemma 8 and the fact that the c_m are non-decreasing then gives

$$\begin{aligned}
\sum_{j=1}^{r} p_{i_j} cc_{m_j} &\leq cw(v) \left[\sum_{j=1}^{r-1} c_{m_j} \left(3 \sum_{m=m_j}^{m_{j+1}-1} 2^{-cc_m} \right) + 2c_{m_r} \sum_{m=m_r}^{\infty} 2^{-cc_m} \right] \\
&\leq 3cw(v) \sum_{m=m_1}^{\infty} c_m 2^{-cc_m} \\
&\leq 3cw(v) \sum_{m \geq m_\epsilon}^{\infty} c_m 2^{-cc_m}.
\end{aligned}$$

21

We are given that $\sum_{m=1}^{\infty} c_m 2^{-cc_m}$ converges. Thus $g(m_\epsilon) \downarrow 0$ as $m_\epsilon \to \infty$ where $g(x) = \sum_{m \geq x}^{\infty} c_m 2^{-cc_m}$.

Note that as N_ϵ increases, m_ϵ increases. Given ϵ, we now choose N_ϵ to be the smallest value such that $g(m_\epsilon) \leq \frac{\epsilon}{6}$. Note that N_ϵ is independent of v.

Combine the above bounds:

$$
\begin{aligned}
\sum_{\substack{v \in N_T}} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)|=1}} w(v) E^*(v,m) \;=\; & \sum_{\substack{v \in N_T}} \sum_{\substack{1 \leq m \leq m_\epsilon \\ |I_m^*(v)|=1}} w(v) E^*(v,m) + \sum_{\substack{v \in N_T}} \sum_{\substack{m_\epsilon < m \\ |I_m^*(v)|=1}} w(v) E^*(v,m) \\
\leq \; & cN_\epsilon + \sum_{v \in N_T} \frac{\epsilon}{2} cw(v)
\end{aligned}
$$

Recall from Lemma 1 and the fact that $\forall m, c_m \geq 1$,

$$
C(T) = \sum_{v \in N_T} \sum_{m=1}^{t} c_m \cdot w_m(v) \geq \sum_{v \in N_T} \sum_{m=1}^{t} w_m(v) = \sum_{v \in N_T} w(v).
$$

Thus, we have just seen that

$$
\sum_{\substack{v \in N_T}} \sum_{\substack{1 \leq m \leq t \\ |I_m^*(v)|=1}} w(v) E^*(v,m) \leq cN_\epsilon + \frac{\epsilon}{2} cC(T).
$$

Plugging back into Corollary 7 gives

$$
cC(T) - H(p_1, \ldots, p_n) \leq 2(1-p_1) + c(c_2 - c_1) + cN_\epsilon + \frac{\epsilon}{2} cC(T)
$$

which can be rewritten as

$$
C(T) - \frac{1}{1-\frac{\epsilon}{2}} \frac{1}{c} H(p_1, \ldots, p_n) \leq \frac{1}{1-\frac{\epsilon}{2}} \frac{1}{c} \left(2(1-p_1) + c(c_2 - c_1) + cN_\epsilon \right)
$$

We may assume that $\epsilon \leq 1/2$, so $1 + \epsilon \geq \frac{1}{1-\frac{\epsilon}{2}}$. Thus

$$
C(T) - (1+\epsilon)\frac{1}{c} H(p_1, \ldots, p_n) \leq f(\mathcal{C}, \epsilon)
$$

where

$$
f(\mathcal{C}, \epsilon) = \frac{4}{3}\left(\frac{2}{c} + (c_2 - c_1) + N_\epsilon \right). \tag{9}
$$

This can then be rewritten as

$$
\begin{aligned}
R = C(T) - \frac{1}{c} H(p_1, \ldots, p_n) \;\leq\; & \epsilon \frac{1}{c} H(p_1, \ldots, p_n) + f(\mathcal{C}, \epsilon) \\
\leq\; & \epsilon OPT + f(\mathcal{C}, \epsilon)
\end{aligned}
$$

proving the Theorem. $\qquad \square$

6 Examples

We now examine some of the bounds derived in the last section and show how they compare to the old bound of $(1 - p_1 - p_n) + cc_t$ stated in (3). In particular, we show that for large families of costs the old bounds go to infinity while the new ones give uniformly constant bounds.

<u>Case 1:</u> $\mathcal{C}_\alpha = (c_1, c_2, \ldots, c_{t-1}, \alpha)$ with $\alpha \uparrow \infty$.
We assume $t \geq 3$ and all of the c_i, $i < t$, are fixed. Let $c^{(\alpha)}$ be the root of the corresponding characteristic equation $1 = 2^{-c\alpha} + \sum_{i=1}^{t-1} c^{-cc_i}$. Note that $c^{(\alpha)} \downarrow \bar{c}$ where $\bar{c}$ is the root of $1 = \sum_{i=1}^{t-1} c^{-cc_i}$. Let (NR_α) R_α be the (normalized) redundancy corresponding to $\mathcal{C}_\alpha$.

For any fixed α, the old bound (3) would give

$$NR_\alpha \leq (1 - p_1 - p_n) + c^{(\alpha)}\alpha, \quad R_\alpha \leq \frac{(1 - p_1 - p_n)}{c^{(\alpha)}} + \alpha,$$

the right hand sides of both of which tend to ∞ as α increases. Compare this to Theorem 3 which gives a uniform bound of

$$
\begin{aligned}
NR_\alpha &\leq 2(1 - p_1) + \max\left(c^{(\alpha)}(c_2 - c_1), 1 + \log t\right) \\
&\leq 2(1 - p_1) + \max\left(c^{(c_{t-1})}(c_2 - c_1), 1 + \log t\right)
\end{aligned}
$$

and

$$R_\alpha \leq \frac{NR_\alpha}{c^{(\alpha)}} \leq \frac{2(1 - p_1) + \max\left(c^{(c_{t-1})}(c_2 - c_1), 1 + \log t\right)}{\bar{c}}.$$

For concreteness, we examine a special case of the above.

Example 1 *Let $t = 3$ with $c_1 = c_2 = 1$ and $c_3 = \alpha \geq 1$. The old bounds (3) gives an asymptotically infinite error as $\alpha \to \infty$. The bound from Theorem 3 is*

$$NR_\alpha \leq 2(1 - p_1) + \max\left(c^{(\alpha)}(c_2 - c_1), 1 + \log t\right) \leq 3 + \log 3$$

independent of α. Since $c^{(\alpha)} \geq \bar{c} = 1$ we also get

$$R_\alpha = \frac{NR_\alpha}{c^{(\alpha)}} \leq 3 + \log 3.$$

<u>Case 2:</u> A finite alphabet that approaches an infinite one.
Let $\mathcal{C}$ be an infinite sequence of letter costs such that there exists a $K > 0$ satisfying for all j, $d_j = |\{i \mid j \leq c_i < j\}| \leq K$. Let c be the root of

the characteristic equation $1 = \sum_{i=1}^{\infty} 2^{-cc_i}$. Let $\Sigma^{(t)} = \{\sigma_1, \ldots, \sigma_t\}$ and its associated letter costs be $\mathcal{C}^{(t)} = \{c_1, \ldots, c_t\}$. Let $c^{(t)}$ be the root of the corresponding characteristic equation $1 = \sum_{i=1}^{t} 2^{-cc_i}$ and (NR_t) R_t be the associated (normalized) redundancy. Note that $c^{(t)} \uparrow c$ as t increases.

For any fixed t, the old bound (3) would be $NR_t \le (1 - p_1 - p_n) + c^{(t)} c_t$ which goes to ∞ as t increases. Lemma 9 tells us that

$$\beta^{(t)} = \max_{1 \le m \le t} 2^{c^{(t)} c_m} \sum_{i=1}^{t} 2^{c^{(t)} c_i} \le \frac{2^c K}{1 - 2^{-c^{(t)}}} \le \frac{2^c K}{1 - 2^{-c^{(2)}}}.$$

so, from Theorem 2 and the fact that $\forall t$, $c^{(2)} \le c^{(t)} < c$, we get

$$NR_t \le 2(1 - p_1) + \max\left(c(c_2 - c_1),\ 1 + c + \log \frac{K}{1 - 2^{-c^{(2)}}} \right).$$

Note that if all of the c_m are integers, then the additive factor c will vanish.

Example 2 *Let $\mathcal{C} = (1, 2, 3, \ldots)$. i.e., $c_m = m$. The old bounds (3) gives an asymptotically infinite error as $\alpha \to \infty$.*

For this case $c = 1$ and $K = 1$. $c^{(2)}$ is the root of the characteristic equation $1 = 2^{-2} + 2^{-2c}$. Solving gives $2^{-c^{(2)}} = \frac{\sqrt{5}-1}{2}$ and $c^{(2)} = 1 - \log(\sqrt{5} - 1) \approx 0.694 \ldots$. Plugging into our equations gives

$$
\begin{aligned}
NR_t &\le 2(1 - p_1) + \max\left(c(c_2 - c_1),\ 1 + \log\left(\frac{K}{1 - 2^{-c^{(2)}}} \right) \right) \\
&= 2(1 - p_1) + 1 + \log\left(\frac{2}{3 - \sqrt{5}} \right) \le 4.388
\end{aligned}
$$

and

$$R_t = \frac{NR_t}{c^{(t)}} \le \frac{NR_t}{c^{(2)}} \le 6.232.$$

Case 3: An infinite case when $d_j = O(1)$.
In this case just apply Lemma 9 directly.

Example 3 *Let $\mathcal{C}$ contain d copies each of $i = 1, 2, 3, \ldots$, i.e., $c_m = 1 + \lfloor \frac{m-1}{d} \rfloor$. Note that $K = d$. If $d = 1$, i.e., $c_m = m$, then $c = K = 1$ and*

$$R = NR \le 2(1 - p_1) + 2.$$

If $d > 1$ then $A(x) = \sum_{m=1}^{\infty} c_m z^m = \frac{dz}{1-z}$. The solution α to $A(\alpha) = 1$ is $\alpha = \frac{1}{d+1}$, so $c = -\log \alpha = \log(d + 1)$. The lemma gives

$$NR \le 2(1 - p_1) + \left(1 + \log\left(\frac{K}{1 - 2^{-c}} \right) \right) \le 3 + \log(d+1), \quad R \le 1 + \frac{3}{\log(d + 1)}.$$

<u>Case 4:</u> d_j are integral and satisfy a linear recurrence relation.

In this case the generating function $A(z) = \sum_{j=1}^{\infty} d_j z^j = \sum_{m=1}^{\infty} z^{c_m}$ can be written as $A(z) = \frac{P(z)}{Q(z)}$ where $P(z)$ and $Q(z)$ are relatively prime polynomials. Let γ be a smallest modulus root of $Q(z)$. If γ is the unique root of that modulus (which happens in most interesting cases) then it is known that $d_j = \Theta(j^{d-1}\gamma^{-j})$ (which will also imply that γ is positive real) where d is the multiplicity of the root. There must then exist some $\alpha < \gamma$ such that $A(\alpha) = 1$. By definition $c = -\log \alpha$. Furthermore, since $\alpha < \gamma$ we must have that $\sum_{j=1}^{\infty} d_j j \alpha^j = \sum_{m=1}^{\infty} c_m \alpha^{c_m}$ also converges, so Theorem 4 applies.

Note that

$$h(x) = \sum_{j=x}^{\infty} d_j j \alpha^j = O\left(\sum_{j=x}^{\infty} j^{d-1} j \left(\frac{\alpha}{\gamma} \right)^j \right) = O\left(x^d \left(\frac{\alpha}{\gamma} \right)^x \right),$$

implying

$$h^{-1}(\epsilon) = \log_{\gamma/\alpha} 1/\epsilon + O(\log \log 1/\epsilon)$$

where we define

$$h^{-1}(\epsilon) = \max\{x \mid h(x) \le \epsilon, \ h(x-1) > \epsilon\}.$$

Working through the proof of Theorem 4 we find that when the c_m are all integral,

$$\forall m', \quad g(m') = \sum_{m \ge m'} c_m 2^{-cc_m} \quad = \quad \sum_{m \ge m'} c_m \alpha^{c_m}$$

$$\le \quad \sum_{j \ge c_{m'}} j d_j \alpha_j = h(c_{m'}).$$

Recall that $m_\epsilon = \max\{m \mid c_m \le N_\epsilon\}$. Then $g(m_\epsilon) \le h(N_\epsilon)$. Since $g(m_\epsilon) \le \epsilon/6$,

$$N_\epsilon \le h^{-1}(\epsilon/6) = \log_{\gamma/\alpha} 1/\epsilon + O(\log \log 1/\epsilon)$$

and thus our algorithm creates a code T satisfying

$$C(T) - OPT \le \epsilon OPT + \log_{\gamma/\alpha} 1/\epsilon + O(\log \log 1/\epsilon). \qquad (10)$$

Example 4 *Consider the case where $d_j = F_j$, the jth Fibonacci number, $F_1 = 1$, $F_2 = 1$, $F_3 = 2$,.... It's well known that $A(z) = \sum_{j=1}^{\infty} d_j z^j = \frac{x}{1-x-x^2}$ and $F_j = \frac{\phi^n - (1-\phi)^n}{\sqrt{5}}$ where $\phi = \frac{1+\sqrt{5}}{2}$. Thus $d_j = \frac{\gamma^{-j}}{\sqrt{5}} + e_n$ where $\gamma = \phi^{-1}$ and $|e_n| < 1$. Solving $A(\alpha) = 1$ gives $\alpha = \sqrt{2} - 1 \approx .4142\dots$ (and $c = -\log \alpha = 1.2715\dots$). (10) gives a bound on the cost of the redundancy of our code with $\frac{\gamma}{\alpha} = \frac{2}{(1+\sqrt{5})(\sqrt{2}-1)} \approx 1.492\dots$.*

<u>Case 5:</u> An example for which there is no known bound.

An interesting open question is how to bound the redundancy for the case of balanced words described at the end of Section 3. Recall that this had d_j integral with $d_j = 0$ for $j = 0$ and odd j and for even $j > 0$, $d_j = 2C_{j/2-1}$ where $C_i = \frac{1}{i+1}\binom{2i}{i}$ is the ith *Catalan number*. It's well known that $\sum_{j=0}^{\infty} C_j x^j = \frac{1}{2x}(1 - \sqrt{1-4x})$ so

$$A(x) = \sum_{m=1}^{\infty} x^{c_m} = \sum_{j=1}^{\infty} d_j x^j = 1 - \sqrt{1-4x^2}.$$

Solving for $A(\alpha) = 1$ gives $\alpha = \frac{1}{2}$ and $c = -\log\alpha = 1$. On the other hand,

$$\sum_{m=1}^{\infty} c_m x^{c_m} = \sum_{j=1}^{\infty} j d_j x^j = 2\sum_{j=1}^{\infty} \binom{2(j-1)}{j-1}(x^2)^j = \frac{x^2}{\sqrt{1-4x^2}}$$

so this sum *does not* converge when $x = 1/2$. Thus, we can not use Theorem 4 to bound the redundancy. Some observation shows that this $\mathcal{C}$ does not satisfy any of our other theorems either. It remains an open question as to how to construct a code with "small" redundancy for this problem, i.e., a code with a constant additive approximation or something similar to Theorem 4.

7   Conclusion and Open Questions

We have just seen $O(n \log n)$ time algorithms for constructing almost optimal prefix-free codes for source letters with probabilities $p_1, \ldots, p_n$ when the costs of the letters of the encoding alphabet are unequal values $\mathcal{C} = \{c_1, c_2, \ldots\}$. For many finite encoding alphabets, our algorithms have provably smaller redundancy (error) than previous algorithms given in [20, 8, 23, 2]. Our algorithms also are the first that give provably bounded redundancy for some infinite alphabets.

There are still many open questions left. The first arises by noting that, for the finite case, the previous algorithms were implicitly constructing *alphabetic codes*. Our proof explicitly uses the fact that we are only constructing general codes. It would be interesting to examine whether it is possible to get better bounds for alphabetic codes (or to show that this is not possible).

Another open question concerns Theorem 4 in which we showed that if $\sum_{m=1}^{\infty} c_m 2^{-cc_m} < \infty$, then,

$$\forall \epsilon > 0, \quad C(T) - OPT \le \epsilon OPT + f(\mathcal{C}, \epsilon).$$

Is it possible to improve this for some general $\mathcal{C}$ to get a purely additive error rather than a multiplicative one combined with an additive one?

Finally, in Case 5 of the last section we gave a natural example for which the root c of $\sum_{i=1}^{\infty} 2^{-cc_m} = 1$ exists but for which $\sum_{m=1}^{\infty} c_m 2^{-cc_m} = \infty$ so that we can not apply Theorem 4 and therefore have no error bound. It would be interesting to devise an analysis that would work for such cases as well.

Acknowledgments

References

[1] Julia Abrahams. Code and parse trees for lossless source encoding. *Communications in Information and Systems*, 1(2):113–146, April 2001.

[2] Doris Altenkamp and Kurt Melhorn. Codes: Unequal probabilies, unequal letter costs. *Journal of the Association for Computing Machinery*, 27(3):412–427, July 1980.

[3] Toby Berger and Raymond W. Yeung. Optimum '1' ended binary prefix codes. *IEEE Transactions on Information Theory*, 36(6):1435–1441, 1990.

[4] N. M. Blachman. Minimum cost coding of information. *IRE Transactions on Information Theory*, PGIT-3:139–149, 1954.

[5] P. Bradford, M. Golin, L. L. Larmore, and W. Rytter. Optimal prefix-free codes for unequal letter costs and dynamic programming with the monge property. *Journal of Algorithms*, 42:277–303, 2002.

[6] Renato M. Capocelli, Alfredo De Santis, and Giuseppe Persiano. Binary prefix codes ending in a "1". *IEEE Transactions on Information Theory*, 40(4):1296–1302, 1994.

[7] Norbert Cott. *Characterization and Design of Optimal Prefix Codes*. PhD Thesis, Stanford University, Department of Computer Science, June 1977.

[8] I. Csisz'ar. Simple proofs of some theorems on noiseless channels. *Inform. Contr.*, 514:285–298, 1969.

[9] E. N. Gilbert. How good is morse code. *Inform Control*, 14:585–565, 1969.

[10] E. N. Gilbert. Coding with digits of unequal costs. *IEEE Trans. Inform. Theory*, 41:596–600, 1995.

[11] M. Golin and G. Rote. A dynamic programming algorithm for constructing optimal prefix-free codes for unequal letter costs. *IEEE Transactions on Information Theory*, 44(5):1770–1781, 1998.

[12] Mordecai J. Golin, Claire Kenyon, and Neal E. Young. Huffman coding with unequal letter costs. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC'02)*, pages 785–791, 2002.

[13] K. Hinderer. On dichotomous search with direction-dependent costs for a uniformly hidden objec. *Optimization*, 21(2):215–229, 1990.

[14] D. A. Huffman. A method for the construction of minimum redundancy codes. In *Proc. IRE 40*, volume 10, pages 1098–1101, September 1952.

[15] K. A. S. Immink. *Codes for Mass Data Storage Systems*. Shannon Foundations Publishers, 1999.

[16] I. Itai. Optimal alphabetic trees. *SIAM J. Computing*, 5:9–18, 1976.

[17] Sanjiv Kapoor and Edward M. Reingold. Optimum lopsided binary trees. *Journal of the Association for Computing Machinery*, 36(3):573–590, July 1989.

[18] Richard Karp. Minimum-redundancy coding for the discrete noiseless channel. *IRE Transactions on Information Theory*, IT-7:27–39, January 1961.

[19] Donald E. Knuth. *The Art of Computer Programming, Volume III: Sorting and Searching*. Addison-Wesley, 1973.

[20] R. M. Krause. Channels which transmit letters of unequal duration. *Inform. Contr.*, 5:13–24, 1962.

[21] Jia-Yu Lin, Ying Liu, and Ke-Chu Yi. Balance of 0, 1 bits for huffman and reversible variable-length coding. *IEEE Transactions on Communications*, 52(3):359–361, 2004.

[22] R.S. Marcus. *Discrete Noiseless Coding*. M.S. Thesis, MIT E.E. Dept, 1957.

[23] K. Mehlhorn. An efficient algorithm for constructing nearly optimal prefix codes. *IEEE Trans. Inform. Theory*, 26:513–517, September 1980.

[24] Yale N. Patt. Variable length tree structures having minimum average search time. *Commun. ACM*, 12(2):72–76, 1969.

[25] Robert Sedgewick and Philippe Flajolet. *An introduction to the analysis of algorithms*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1996.

[26] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[27] L. E. Stanfel. Tree structures for optimal searching. *Journal of the Association for Computing Machinery*, 17(3):508–517, July 1970.

[28] B. Varn. Optimal variable length codes (arbitrary symbol cost and equal code word probability). *Information Control*, 19:289–301, 1971.